DOCUMENTATION
INCORPORATED

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

THE PROBLEM OF MACHINE ASSOCIATION OF IDEAS

TECHNICAL REPORT NO. 1

PREPARED UNDER

CONTRACT NO. Nonr-1305(00)

for

THE OFFICE OF NAVAL RESEARCH

NOVEMBER
1953

DOCUMENTATION
Incorporated

WASHINGTON 8, D. C.

DOCUMENTATION
INCORPORATED

9521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

I. Introduction

The work of Documentation Incorporated in the logical analysis of information systems has led to the threshold of a wholly new development in the organization and utilization of information systems, especially such systems as are the basis of intelligence activities and military operations. We now see the possibility of introducing into any system of information not only a method for the rapid and comprehensive indexing of such information, but also a means of searching the system from any point of entry for all the ideas in it which are factually associated with one another.

The significance of this possibility can best be indicated by the following statement taken from Dr. Vannevar Bush's prophetic article, "As We May Think"[1]:

> The real heart of the matter of selection, however, goes deeper than a lag in the adoption of mechanisms by libraries, or a lack of development of devices for their use. Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path. The human mind does not work that way. It operates by association ... Man cannot hope

[1] Bush, V.. As We May Think. Atlantic Monthly, July, 1945.

> fully to duplicate this mental process artificially, but he certainly ought to be able to learn from it. In minor ways he may even improve, for his records have relative permanency. The first idea, however, to be drawn from the analogy concerns selection. Selection by association, rather than by indexing, may yet be mechanized. One cannot hope thus to equal the speed and flexibility with which the mind follows an associative trail, but it should be possible to beat the mind decisively in regard to the permanence and clarity of the items resurrected from storage.

In organizing a body of information we need be able not only to recover material indexed under any heading by searching under that heading, but also to trace all the ideas associated with the heading with which we start through the total system. This involves what Charles Bernier has called the "coincidence of pertinent vocabularies" He says:

> "In my opinion, the problem of bringing pertinent vocabularies into coincidence is the crucial one for all documentation systems whether operated by classifications, indexes, or machines. The system must display its vocabulary for selection. For the field of chemistry, this vocabulary is so enormous and is growing so rapidly that is impossible because of limitations of time and eyesight to read all of the terms for the purpose of making the proper and complete selection. To facilitate selection, indexes give users cross references and display suggestive and new terms in the modifications under headings. The searchers aid in bringing their vocabularies into coincidence by courses of study, textbooks, reference works, reviews, monographs, dictionaries, current literature, indexes, and the like.

At present, the process of bringing vocabularies into coincidence for broad, generic searches is probably the slowest of all processes in the use of documentation systems. The important point to note is that all documentation systems, including mechanized ones, are faced with this problem of bringing pertinent vocabularies into coincidence and that if coincidence is not affected then significant information is certain to be missed. Difficulties with broad generic searches, formerly attributed by some of us interested in mechanization solely to indexes, are, in reality, caused by the difficulties of bringing vocabularies into coincidence."[2]

The tracing of such "coincidences" or associations is what is usually meant by browsing. In card catalogs and book indexes such browsing is possible, but haphazard. In machine indexing systems as hitherto envisioned, it has not even been possible.

Suppose in studying a landing operation at Beach X, we began with ideas relating to "landing craft," "off-shore support," and "distance from base." In an ordinary index we could look up information under any one or all of these headings in turn; in a coordinate index as will be explained later, we could go further and search for information having specific reference to the combination or logical conjunction of the three headings.

What is proposed here is the construction of a system of associations whereby starting with a few ideas there will be disclosed to the searcher all other ideas in the system associated with those with which he begins his search. That is, beginning with the three ideas above, the system

[2] Charles Bernier, "Organizing Abstract Information", unpublished paper read before American Documentation Institute, November 6, 1953.

would disclose the associated ideas of supply, tide, meteorological conditions, geological formation of beach, slope of beach, etc. It is believed that tracing such associations through a system of ideas can be mechanized, and that this will introduce a radically shortened time dimension and assurance of completeness in the analysis and evaluation of any information search or intelligence or operational situation.

The importance of this development for operations research, which has been defined as "related research", cannot be over-emphasized. In more narrow and restricted fields of inquiry, the retrieval of stored information by searching under specified discreet index headings may be adequate; but the very nature of operational research makes it both desirable and necessary to be able to range freely and rapidly through sets of associated ideas.

In order to clarify fully the significance of this possibility and its promise of making possible for the first time the truly economic and efficient machine manipulation of a total system of ideas in intelligence and operational analysis, it is necessary to present an account of the developments in both theory and practice which have led up to it.

Uniterm System of Coordinate Indexing

The Uniterm system of coordinate indexing was developed by Documentation Incorporated under a contract with a Department of Defense agency, and applications of this system are now being used in a number of government and commercial agencies, including the Office of Naval Research. It was the successful application of logical analytical methods to the problem of size, completeness, and efficiency of any index which indicated the possibility of using these methods to go beyond the indexing of information and into the hitherto unexplored field of the mechanical association of ideas. No machine now available can manipulate efficiently the number of actual distinct items of information in any large technical or intelligence library. But if logical analysis is first employed to reduce such items of information to the relatively small set of unit ideas, their manipulation - their association, dis-association and reassociation by machines - becomes a practical and exciting possibility.

The usual indexing in previous systems is by means of a combination of words which can be arranged internally many different ways and, consequently, requires listing in many different places. In principle, any two-word index heading can be arranged and listed in two ways; any three-word heading requires for completeness six arrangements and six listings; any four-word heading, 24 permutations; any five-word heading, 120. The mathematical law which describes the total possible arrangements and listings is "n factorial".

The most generally employed method of limiting the size of any index is to limit by convention the permutations (cross references or multiple filings) of the words in the headings. However necessary as practical measures, the conventions or rules which limit the size of any given index also limit its usefulness. Hence, any indexer using standard methods must seek an acceptable compromise between cost and utility.

The Uniterm system of coordinate indexing resolves this dilemma by analyzing the subject material into a basic vocabulary and providing means for coordinating the elements of the vocabulary into any specific combination desired. No term in the system is repeated more than once regardless of the number of other terms with which it may be combined or the order of such combinations. Through this method, the indexing apparatus to any collection of reference materials can be reduced from 50 to 90 per cent in bulk and made more efficient as bibliographical and reference tools.

The technical advantages of the Uniterm system can be summarized as follows:

1. Every term in the Uniterm system is a filing term or access point.
2. Since there are no subdivisions, every term in the Uniterm system is on equal footing with every other and can be the subject of a complete search. It is impossible in any standard index to make a complete search of reference material under any heading which has been used as a subdivision.
3. All "see" references which are required in a standard system by virtue of the order of words in index headings are eliminated.
4. All "see also" references from general to specific subjects are eliminated.
5. The subjective choice of the indexer between possible permutations of multiple-term descriptions is eliminated.
6. Since every item in the Uniterm system is a filing word and each term in the system appears only once as a filing word, any need for searching for the "proper subdivision" in the proper phrase is eliminated.

II. The Logic of Association of Ideas and its Mechanization

When a document is indexed under the Uniterm System the subject material is analyzed in the basic units of language which are required to retrieve the subject content. Each Uniterm represents one of the ideas about which there is positive information in the document. The total number of terms by which the document is indexed represents the positive associations of these basic ideas or in other terms "the positive Boolean conjunctive functions."

As a large system of information is analyzed into Uniterms, there will be many terms generated from the documents of the system. Since some of the reports or other documents will be on the same subjects or otherwise related to one another, their Uniterms or unit ideas will themselves be associated. However, not all of the Uniterms in the vocabulary of the system will be related nor will all the possible combinations of Uniterms yield positive valid functions.

Suppose, for example, that in a particular coordinate index, information is stored under the following terms:

carriers
range
fire power
fuel supply
railroads

In such a system, there might also be valid information under the combination, "carriers - range - fire power - fuel supply," and under the combination "railroads - fuel supply - range." But it is not likely that there would be information under the combinations "railroads - fuel supply - fire power" or "carriers - range - Railroads." The combinations - or associations of ideas - which yield valid information can be considered "positive Boolean conjunctive functions," whereas these combinations which are possible in any system but which do not yield information can be considered empty functions.

In any system of "n" terms, the total number of conjunctive functions or possible associations of ideas is equal to $2^n - 1$. It can be seen that with a system of a thousand terms or more the number of possible associations or conjunctive functions is very, very large. On the other hand, the number of associations which yield information will only be a tiny fraction of the total number of possible functions or associations of ideas. For example, suppose in any system of a thousand ideas, no association yielding information contained more than ten ideas, it would follow that adding the 11th idea to any combination of 10 would result in any empty function or an association of ideas which would not lead to information. Therefore, the positive associations are small fraction of the possible associations in the system. This means, further, that the machine searching of any system of ideas for the positive associations can be a rapid self-limiting and self-converging process.

Dictionary of Associations

There are many ways to record conjunctive functions or associations of ideas ranging from simple notched or punched cards to complicated electronic computing machines. Part of our work in the next year will consist in deciding on the best mechanism to trace any association of ideas through a properly organized system of information. The instrument for such a search would be, in effect, a mechanized dictionary of associated ideas.

In order to see how such a dictionary might operate, consider the following simple example taken from Philip Morse's "Methods of Operations Research." Morse points out that in discussions of the desirability of providing merchant ships with antiaircraft guns, the decision seemed to rest on the number of planes shot down by armed merchant ships. According to Morse, "It took an operations research worker to point out that, even though the enemy planes were not shot down, the antiaircraft guns were valuable because they decreased the accuracy of the enemy planes enough to lessen the chance that the merchant vessel be sunk.[3] One would expect that if the relevant information had been properly indexed or associated in the dictionary, a search of the dictionary would have yielded the necessary additional parameters or associated ideas - beginning with any one of the ideas, namely, "accuracy of bombing" and "ratio of unarmed ships to armed ships sunk by enemy aircraft."

Such a system of associations would reject not only empty functions but all associations which are only partial expressions of more complete and specific associations. For example, beginning with the terms "A" "D",

[3] Morse, Philip M. and Kimball, George: Methods of Operations Research, 1951, p.6.

the dictionary might disclose:

A D F

A D M N P

A D M O R S

A D O R S T U

It can be seen from the example above the "A D M" is a partial association of the two more complete functions or associations "A D M N P", "A D M O R S". By using the relation of logical alternation in the Uniterm system, we can create the association "A D M and (NP or ORS)". Going still further, by using negative relations, e.g., the requirement that "R" not be associated in the desired complex, we eliminate "A D M O R S" and are left with a single association "A D M N P".

In the example used above, the ideas associated with AD are presented in tabular form as might be the case with a printed dictionary. However, a printed dictionary of associations for even a medium sized system of information would be too large and too complicated for convenient use. For example, a system of 5000 terms used to analyze 50,000 documents might generate close to 500,000[4] different positive associations. Whereas this is a small number as compared with the billions and billions of possible functions, it is still too large to be displayed in a standard printed book. What is required is a method of showing only the basic vocabulary, and a method for finding and displaying the positive functions of the basic vocabulary on demand.

[4] Our next report will contain an account of our method of arriving at a figure for the maximum and minimum number of positive functions in any system of information.

Thus, the basic concept presented here is that of an externalized permanent memory in the form of a mechanised dictionary of associations or "pertinent coincidences" of all the trails of ideas which have been put into the system by many people. This associative memory then does not depend on having these people available for a problem and does not lose information when personnel leave. Of course putting the ideas into the machine properly presents many problems which will be studied.

The Mechanical Dictionary will not create associations which have not been previously recorded any more than browsing in an index results in the discovery of unindexed combinations. But it will, as Dr. Bush suggests "beat the mind decisively in regard to the permanence and clarity of the items ressurected from storage."

III. Program to be Followed

In order to get an idea of the magnitude of the problem of generating the Dictionary of Associations, we are currently examining the second level associations in a Uniterm Index to a part of the Armed Services Technical Information Agency collection. This is a collection which embraces all fields of technology and is representative of a broad range of associations. We are also studying the second level associations in a collection being compiled under the ONR-BuAer program for aircraft instrumentation. This includes the relatively narrow field of instrumentation, data presentation and human engineering as applied to airborne equipment. Thus, the ideas in this collection should be inherently associated. The results of these investigations will be presented in the next report.

From the empirical studies of an actual set of associations in these fields, it seems possible to determine the upper and lower limits for the number of positive functions in an actual system.

When the number of associations has been established, it will be possible to study the problems involved in setting such a system into operation in any given collection. It will also be possible to determine the magnitude of the problems of mechanization and presentation of the associations.